Geographic Information in Small Area Estimation: Small Area Models and Spatially Correlated Random Area Effects

TATISTICS IN TRANSITION, December 2005
Vol. 7, No. 3, pp. 609—623

GEOGRAPHIC INFORMATION IN SMALL AREA ESTIMATION: SMALL AREA MODELS AND SPATIALLY CORRELATED RANDOM AREA EFFECTS

A. Petrucci[1], **M. Pratesi, N. Salvati**[2]

ABSTRACT

This work applies the Fay-Herriot model in which spatial information is introduced as auxiliary variables, and generalizes the model by introducing spatially correlated random area effects modelled through the Simultaneously Autoregressive (SAR) process.

The traditional Empirical Best Linear Unbiased Predictor (EBLUP) takes advantage of the between small area-variation. The evidence is that the EBLUP estimator is significantly better than the sample-size dependent estimators, especially when the between small area-variation is not large relative to the within small area variation. This suggests that the location of the small areas may also be relevant in modelling the small area parameters and that further improvement in the EBLUP estimator can be gained by including eventual spatial interaction among random area effects.

The properties of the proposed estimators are evaluated by applying them to two agro-environmental case studies.

***Key words*:** Spatial information, EBLUP, Spatial EBLUP, GIS.

1. Introduction

Geographic information and geographical modelling can be valuable tools in describing and understanding many phenomena. It is a matter of fact that both environmental and social economic phenomena have a spatial distribution conditioned by nature and by the action of man. The spatial distribution of a pollution agent in the soil is the result of the geological conformation of the soil as well as man's actions in the construction of roads, houses and factories. The distribution of crops in a region is another example of this combined action: man

[1] Dipartimento di Statistica "G. Parenti",Università di Firenze, e-mail alex@ds.unifi.it

[2] Dipartimento di Statistica e Matematica Applicata all'Economia, Università di Pisa, e-mail: m.pratesi@ec.unipi.it; salvati@ec.unipi.it

and nature work together and the result is the distribution of cultivated land that is scattered over the surface of a region. When we look at a landscape we clearly recognize the effect of this combined action, and we note the evidence of the famous first law of geography: "everything is related to everything else, but near things are more related than distant things" (Tobler 1970). The law is valid also for small geographical areas: close areas are more likely to have similar values of the target parameter than areas which are far from each other. This evidence suggests that an adequate use of geographic information and geographical modelling can help in producing more accurate estimates for small area parameters. If the target is a small area parameter, geographical information is a valid help in order to take advantage from the information of the related areas.

Geographic information in this context is intended to be the whole set of data about the position in space of the small areas and of the units which are located in the area itself. Also, the spatial relations among areas are of great interest: we refer to the geometric properties of contiguity between areas, and to the distances (Euclidean and not) between them. Section 2 is devoted to describing this additional data and to how they can be easily stored and managed by a Geographical Information System.

In our approach the small area parameters are estimated via a model based perspective. The attention is on the Fay-Herriot model and spatial information is introduced as auxiliary variables. The model is generalized by introducing spatially correlated random area effects. Section 3 illustrates the generalized model and the main references for its specification under SAR (Simultaneous Auto Regressive) area level random effects. Our expectation is that geographic information improves the estimators (it helps in reducing their estimated MSE) depending on the strength of the spatial dependence. Alternative specifications of the model which incorporate the spatial information are tested and discussed in two case studies: estimation of soil erosion in 17 zones of the Rathbun lake watershed (USA), and average production of olives per farm in 42 local economy systems (LESs) of the Tuscany region (Italy).The major findings from these cases are described and discussed in Sections 4. In Section 5 some concluding remarks are made.

2. Geographic information

A Geographic Information System (GIS) is an automated information system that is able to compile, store, retrieve, analyze, and display mapped data. In other words GISs are a set of computer hardware and software for analyzing and displaying spatially referenced features (i.e., points, lines, polygons) with non-geographic attributes (i.e., species, age).

This system is commonly used by government, analysts of environment and society, and many others researchers. Its applications include environmental, urban and demographic studies and transportation analysis to mention only a few

of them. GIS, however, is more than a mapping system. What sets it apart from even the most sophisticated mapping system is its power to analyze data and to present the results of that analysis as useful information for decision makers. The purpose of collecting data for a GIS is (a) to make an inventory of a geographically defined area or (b) to examine and quantify the spatial relationships between area units. From this point of view the potentialities of GIS as a tool for the compilation of statistics, in particularly in the field of small area statistics, are large.

Using a GIS, the available data-set for the study can be combined with the relative map of the study area. This makes the following steps possible:

1. the geocoding of the study area allows for the computation of the coordinates of the centroid of each small area, its geometric properties (extension, perimeter, etc.) and the neighbourhood structure;
2. the study variable and the potential explanatory variables can be referred to the centroid of each small area; the result is an irregular lattice (geocoding).

In spatial statistics a simple way to represent the neighbourhood structure is the proximity matrix (**W**). It is a squared matrix where $w_{ij} = 1$ if region *i* is a neighbour of region *j* and 0 otherwise. The most common way to define neighbourhood is contiguity: an area is a neighbour of another if it shares a common edge or border. There are other ways defining **W** for example by creating more elaborate weights as functions of the length of borders (Cliff and Ord, 1981).

As Pfeffermann (2002) notes, small area estimation methods have up to now made almost no use of the work on spatial analysis carried out by statisticians and mathematical geographers (Cressie, 1991). However, the spatial information can be the basis for building a model for the spatial distribution of the study variable by small area of interest.

3. Small area models

The behavior of spatial phenomena is the result of a mixture of both first order and second order effects. First order effects relate to the variation in the mean value of the process in space (a global or large scale trend). Second order effects result from the spatial correlation structure or the spatial dependence in the process (Bailey and Gatrell, 1995); in other words, the tendency for deviations of the process from its mean to follow each other in neighboring sites (local or small scale effects).

The small area model-based estimators are sensitive to the specification of the model, the choice of covariates and the existence of spatially correlated random area specific effects and they can lead to erroneous inference if the assumed models do not provide a good fit of the data (Rao, 2003).

The traditional Fay-Herriot model takes into account the first order variation: the model can consider geographical auxiliary information as covariates in the fixed or/and random part of the model (sub-section 3.1). The second order effects can be managed by extending the Fay-Herriot model to incorporate spatial correlation between the random small area effects modelled through the Simultaneously Autoregressive (SAR) process (sub-section 3.2).

3.1. Fay-Herriot model

In the estimation for small areas the direct survey estimates often have large sampling variability. It is then common to borrow information from related areas through explicit linking models based on random area specific effects that account for between area variations beyond that is explained by auxiliary variables included in the model (Pfeffermann, 2002). Let m be the number of small areas and p the number of covariates. The basic area level model assumes that the $m \times p$ matrix of the area-specific auxiliary variables, $\mathbf{X}^{\mathrm{T}}$, is related to the $m \times 1$ vector of small area parameters of inferential interest ϑ (mean or/and total) as:

$$\vartheta = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} \tag{1}$$

where $\mathbf{Z}$ is a $m \times m$ matrix of known positive constants, $\mathbf{v}$ is the $m \times 1$ vector of independent random area specific effects with zero mean and $m \times m$ covariance matrix $\sigma_u^2\mathbf{I}$ and $\mathbf{I}$ is $m \times m$ identity matrix. The linking model is combined with the sampling model $\hat{\vartheta} = \vartheta + \mathbf{e}$, where $\hat{\vartheta}$ is the $m \times 1$ vector of direct estimator of ϑ and $\mathbf{e}$ is the vector of independent sampling error with mean $\mathbf{0}$ and $m \times m$ diagonal covariance $\boldsymbol{\varphi}$. It is customary to assume that the matrix $\boldsymbol{\varphi}$ is known. The combined model is (Fay and Herriot, 1979):

$$\hat{\vartheta} = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e} \tag{2}$$

and it is a special case of linear mixed model.

Under the model, the Best Linear Unbiased Predictor (BLUP) $\tilde{\vartheta}_i^{\mathrm{fh}}(\sigma_u^2)$ is extensively used to obtain model based indirect estimators of small area parameters ϑ and associated measures of variability. This approach allows a combination of the survey data with other data sources in a synthetic regression fitted using population area-level covariates. We have applied the empirical version (EBLUP) of the $\tilde{\vartheta}_i^{\mathrm{fh}}(\sigma_u^2)$ predictor: details and formulas can be found in Rao (2003, Chapter 7).

A method to take into account spatial information is to include in the model some geographic covariates for each small area by considering data regarding the spatial location (e.g. the centroid coordinates) and/or other auxiliary geographical variables referred to the same area through the use of the Geographic Information

System. We expect that the inclusion of covariates should be able to take into account spatial interaction when this is due to the covariates themselves. In this case it is reasonable to assume that the random small area effects are independent and that the traditional EBLUP is still a valid predictor.

The geographic information can also be inserted in the random part of the Fay-Herriot model. The geographical coordinates of area centroids may be incorporated in the random part of the model defining a $m \times 2$ vector $\mathbf{Z}$ where the first column represents the latitude of each area and the second column is the longitude of each area. The Fay-Herriot model becomes:

$$\hat{\vartheta} = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \begin{bmatrix} z_{11} & z_{12} \\ M & M \\ z_{i1} & z_{i2} \\ M & M \\ z_{m1} & z_{m2} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \mathbf{e} \tag{3}$$

In this case EBLUP is still the best estimator:

$$\tilde{\vartheta}_i^{fh}(\hat{\sigma}_{u_1}^2, \hat{\sigma}_{u_2}^2) = \mathbf{x}_i\hat{\boldsymbol{\beta}} + \mathbf{b}_i^T \begin{bmatrix} \hat{\sigma}_{u_1}^2 & 0 \\ 0 & \hat{\sigma}_{u_2}^2 \end{bmatrix} \mathbf{Z}^T \times \left\{ diag(\varphi_i) + \mathbf{Z} \begin{bmatrix} \hat{\sigma}_{u_1}^2 & 0 \\ 0 & \hat{\sigma}_{u_2}^2 \end{bmatrix} \mathbf{Z}^T \right\}^{-1} (\hat{\vartheta} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{4}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\hat{\vartheta}$, $\hat{\sigma}_{u_1}^2$, $\hat{\sigma}_{u_2}^2$ are the estimators of the variance components $\sigma_{u_1}^2$, $\sigma_{u_2}^2$ and $\mathbf{b}_i^{\mathrm{T}}$ is a 1×2 vector (z_{1i}, z_{2i}) referred to *i-th* small area.

3.2. Spatial EBLUP

The explicit modelling of spatial effects becomes necessary when (1) we have no geographic covariates able to take into account the spatial interaction in the target variable, (2) we have some geographic covariates, but the spatial interaction is so important that the small area random effects are presumably still correlated. In this case, taking advantage from the information of the related areas appears to be the best solution.

A possibility, is to extend the Fay-Herriot model with spatial correlation between the small area random effects modelling through the Simultaneously Autoregressive (SAR) process (Petrucci and Salvati, 2005; Pratesi and Salvati, 2005; Saei and Chambers, 2003).

The assumption of spatial independence is not unchallenged in the literature. Pfeffermann (2002) shows that, with many areas and large cross-sectional correlations, the loss in efficiency from ignoring correlations among areas can be substantial. However, in his study the correlation between pairs of areas is inserted without considering the spatial location of the areas themselves.

Under the SAR model, the Spatial Best Linear Unbiased Predictor (Spatial BLUP) estimator of ϑ_i is:

$$\begin{aligned}\tilde{\vartheta}_i^{S}(\sigma_u^2,\rho) &= \mathbf{x}_i\hat{\boldsymbol{\beta}} + \mathbf{b}_i^{T}\sigma_u^2\left[(\mathbf{I}-\rho\mathbf{W})(\mathbf{I}-\rho\mathbf{W}^{T})\right]^{-1}\mathbf{Z}^{T}\times \\ &\times\left\{\operatorname{diag}(\varphi_i)+\mathbf{Z}\sigma_u^2\left[(\mathbf{I}-\rho\mathbf{W})(\mathbf{I}-\rho\mathbf{W}^{T})\right]^{-1}\mathbf{Z}^{T}\right\}^{-1}(\hat{\boldsymbol{\vartheta}}-\mathbf{X}\hat{\boldsymbol{\beta}})\end{aligned} \tag{5}$$

where ρ is the spatial dependence parameter, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{V}^{-1}\hat{\boldsymbol{\vartheta}}$, $\mathbf{W}$ is the proximity matrix and $\mathbf{b}_i^{T}$ is a $1\times m$ vector $(0,0\ldots0,1,\ldots0)$ with value 1 in the *i-th* position.

The estimator $\tilde{\vartheta}_i^{S}(\sigma_u^2,\rho)$ depends on the unknown variance components σ_u^2 and ρ. Replacing the parameters with asymptotically consistent estimators $\hat{\sigma}_u^2,\hat{\rho}$, a two stage estimator $\tilde{\vartheta}_i^{S}(\hat{\sigma}_u^2,\hat{\rho})$ is obtained and it is called Spatial EBLUP (Petrucci and Salvati, 2005; Pratesi and Salvati, 2005).

The Mean Squared Error (MSE) of $\tilde{\vartheta}_i^{S}(\hat{\sigma}_u^2,\hat{\rho})$ and its estimates are obtained following the results of Kackar and Harville (1984) and Prasad and Rao (1990). In particular, due to the introduction of the extra parameter ρ, the component g_3 of the MSE becomes:

$$\begin{aligned}g_3(\sigma_u^2,\rho) = \operatorname{tr}\Bigg\{&\begin{bmatrix}\mathbf{b}_i^{T}\left(\mathbf{C}^{-1}\mathbf{Z}^{T}\mathbf{V}^{-1}+\sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^{T}\left(-\mathbf{V}^{-1}\mathbf{Z}^{T}\mathbf{C}^{-1}\mathbf{Z}^{T}\mathbf{V}^{-1}\right)\right)\\ \mathbf{b}_i^{T}\left(\mathbf{A}\mathbf{Z}^{T}\mathbf{V}^{-1}+\sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^{T}\left(-\mathbf{V}^{-1}\mathbf{Z}^{T}\mathbf{A}\mathbf{Z}^{T}\mathbf{V}^{-1}\right)\right)\end{bmatrix}\mathbf{V}\times \\ &\times\begin{bmatrix}\mathbf{b}_i^{T}\left(\mathbf{C}^{-1}\mathbf{Z}^{T}\mathbf{V}^{-1}+\sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^{T}\left(-\mathbf{V}^{-1}\mathbf{Z}^{T}\mathbf{C}^{-1}\mathbf{Z}^{T}\mathbf{V}^{-1}\right)\right)\\ \mathbf{b}_i^{T}\left(\mathbf{A}\mathbf{Z}^{T}\mathbf{V}^{-1}+\sigma_u^2\mathbf{C}^{-1}\mathbf{Z}^{T}\left(-\mathbf{V}^{-1}\mathbf{Z}^{T}\mathbf{A}\mathbf{Z}^{T}\mathbf{V}^{-1}\right)\right)\end{bmatrix}^{T}\overline{\mathbf{V}}(\sigma_u^2,\rho)\Bigg\}\end{aligned} \tag{6}$$

with $\mathbf{C} = \left[(\mathbf{I}-\rho\mathbf{W})(\mathbf{I}-\rho\mathbf{W}^{T})\right]$ and $\mathbf{A} = \sigma_u^2\left[-\mathbf{C}^{-1}\left(2\rho\mathbf{W}\mathbf{W}^{T}-2\mathbf{W}\right)\mathbf{C}^{-1}\right]$ and $\overline{\mathbf{V}}(\sigma_u^2,\rho)$ is the asymptotic covariance matrix of σ_u^2 and ρ. The estimated g_3 is obtained replacing σ_u^2 and ρ by estimators $\hat{\sigma}_u^2$ and $\hat{\rho}$ (Pratesi and Salvati, 2005). The variance components σ_u^2 and ρ can be estimated either by Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) methods, assuming normality of the random effects.

4. Case studies

The properties of previous estimators $\tilde{\vartheta}_i^{fh}(\hat{\sigma}_u^2)$, $\tilde{\vartheta}_i^{fh}(\hat{\sigma}_{u_1}^2,\hat{\sigma}_{u_2}^2)$, and $\tilde{\vartheta}_i^{S}(\hat{\sigma}_u^2,\hat{\rho})$ are evaluated by applying them both to the results of the erosion data collected in the Rathbun Lake Watershed in Iowa (Opsomer et al, 2003) and to the sample survey on the Farm Structure (FSS) in Tuscany (Italy) (ISTAT, 2005). The estimators are unbiased and we investigate their performances in terms of their variability

through the average estimated Mean Squared Error: $\overline{\text{mse}} = \sum_{i=1}^{m} \text{mse}_i / m$. For each estimator we report also the decomposition of MSE in the three components g1, g2 and g3, respectively due to the estimation of the random effects (g1), to the estimation of $\boldsymbol{\beta}$ (g2) and to the estimation of the variance components (g3).

The maps showing the spatial distribution of the estimates per small area are illustrated for each estimator and combination of geographic auxiliary variables used.

4.1. The average soil erosion in 17 zones of the Rathbun lake watershed (USA)

4.1.1. Sampling design and data

In 2000 a survey designed to estimate the amount of erosion delivered to the streams in the Rathbun Lake watershed was completed. The watershed, located in southern Iowa (USA), covers more than 365,000 acres (147,715 ha) in six counties and is divided into 61 sub-watersheds. Within each sub-watershed, three 160-acre (64 ha) plots were selected and a sample of 183 units was obtained. (Opsomer *et al*., 2003). Auxiliary data at the sub-watershed level were the land use and the topography that are considered major determinants of the erosion. Data related to these factors were available for the study region in the form of digital elevation and land use classification coverage.

4.1.2. Results

We have estimated the average erosion per 160-acre plot measured in tons in 17 small areas, resulting by grouping sub-watersheds to a higher hierarchical hydrological level. The small area estimates have been obtained using the following models:

a) EBLUP ($\tilde{\vartheta}_i^{fh}(\hat{\sigma}_u^2)$) with land use and digital elevation;
b) EBLUP ($\tilde{\vartheta}_i^{fh}(\hat{\sigma}_u^2)$) with geographical coordinates (in the UTM coordinate system) of the centroid of each small area (hydrological unit) plus land use and digital elevation;
c) EBLUP ($\tilde{\vartheta}_i^{fh}(\hat{\sigma}_{u_1}^2, \hat{\sigma}_{u_2}^2)$) with land use and digital elevation plus the geographical coordinates of the centroid of each small area (hydrological unit) in the random part of the model;
d) Spatial EBLUP ($\tilde{\vartheta}_i^{S}(\hat{\sigma}_u^2, \hat{\rho})$) with land use and digital elevation.

The neighbourhood structure **W** is defined as follows: spatial weight, w_{ij}, is 1 if area *i* shares an edge with area *j* and 0 otherwise. Sampling variances, φ_i, are estimated smoothing the sampling error associated with the population level estimator (Rao, 1998). The estimated variance, $\hat{\varphi}_i$, is then treated as a proxy to

φ_i. Figure 1 displays the maps of the Rathbun Lake Watershed with models (a-b-c) and (d) Spatial EBLUP estimates for average erosion per 160-acre plot in 17 small areas.

Table 1 shows the correlations between the estimates obtained under models (a), (b), (c), (d): it can be noted that all the correlations are high and positive with the exception of model (c) EBLUP plus geographic coordinates in the random part of the model. In this case the correlation with (a) EBLUP estimator and (d) SEBLUP estimator results are smaller. The introduction of spatial interaction in the random part of the model brings a clearer grouping in the 17 small areas.

Table 1. Correlation matrix between the small area estimates of the average erosion per 160-acre plot

	$\tilde{\vartheta}_i^{fh}(\hat{\sigma}_u^2)$	$\tilde{\vartheta}_i^{fh}(\hat{\sigma}_u^2)$ +*geographic coordinates*	$\tilde{\vartheta}_i^{fh}(\hat{\sigma}_{u_1}^2, \hat{\sigma}_{u_2}^2)$	$\tilde{\vartheta}_i^{S}(\hat{\sigma}_u^2, \hat{\rho})$
$\tilde{\vartheta}_i^{fh}(\hat{\sigma}_u^2)$	1			
$\tilde{\vartheta}_i^{fh}(\hat{\sigma}_u^2)$ +*geographic coordinates*	0.84	1		
$\tilde{\vartheta}_i^{fh}(\hat{\sigma}_{u_1}^2, \hat{\sigma}_{u_2}^2)$	0.62	0.91	1	
$\tilde{\vartheta}_i^{S}(\hat{\sigma}_u^2, \hat{\rho})$	0.97	0.92	0.73	1

Table 2 reports the average estimated MSE per plot and its decomposition in g_1, g_2 and g_3 for the EBLUP and Spatial EBLUP estimators. The EBLUP predictor under model (c), with geographic coordinates in the random part of the model, provides estimates with the smallest average estimated mean squared error. The model performs better than the Spatial EBLUP estimator. It is probably due to the explanatory power of geographic coordinates which are sufficient to take into account the existing spatial interaction. Under model (d) the estimated spatial autoregressive coefficient suggests a moderate spatial relationship: $\hat{\rho}$ is 0.417 (s.e.=0.439). In fact, a simulation study, carried out to assess the accuracy of the Spatial EBLUP estimator, shows that the gain in relative efficiency of the Spatial EBLUP is relevant especially where the spatial correlation is high (Pratesi and Salvati, 2005). The simulation experiment provides evidence on the design bias of the estimators: they are both slightly design biased; when the spatial correlation increases, the design bias of Spatial EBLUP seems to decrease more than the traditional EBLUP.

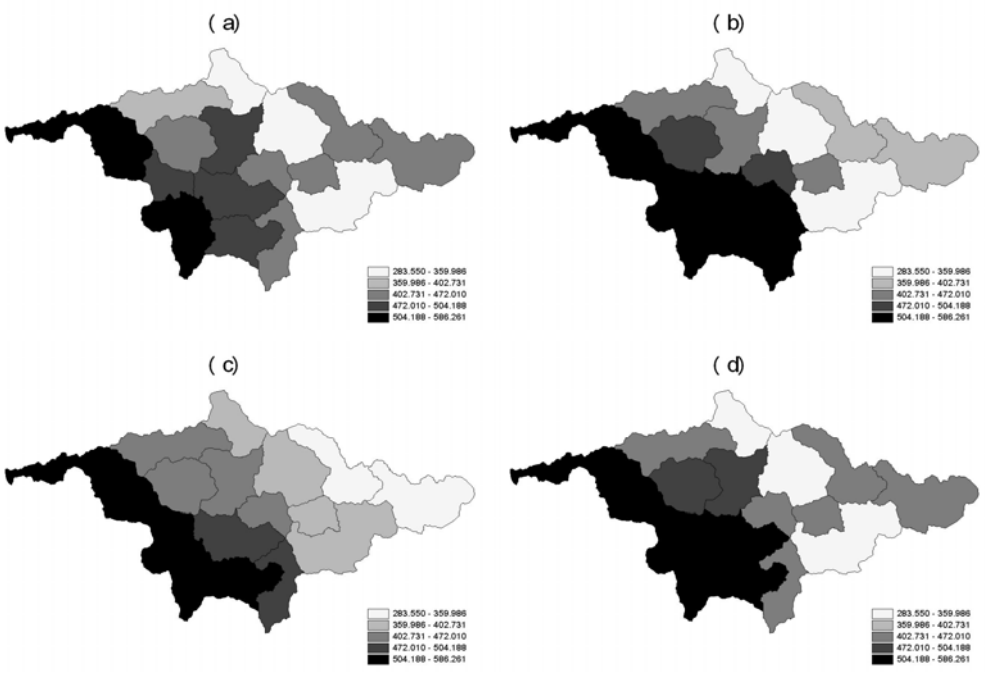
In Table 2 the three estimated components of MSE are ranked in the same order for EBLUP and SEBLUP with the exception of model (c) where g_2, the variability due to the estimation of $\boldsymbol{\beta}$, is the largest one. Note that g_3 for model (d)

is higher than the same component in the EBLUP models (a), (b) and (c) due to the estimation of the additional parameter ρ.

Table 2. Average Estimated MSE of EBLUP and Spatial EBLUP estimators.

Estimator	mse	g_1	g_2	g_3
a) $\tilde{\vartheta}_i^{fh}(\hat{\sigma}_u^2)$	5021.993	3734.116	680.319	284.183
b) $\tilde{\vartheta}_i^{fh}(\hat{\sigma}_u^2)$ +*geographic coordinates*	4000.232	1980.931	1312.091	317.503
c) $\tilde{\vartheta}_i^{fh}(\hat{\sigma}_{u1}^2, \hat{\sigma}_{u2}^2)$	1602.442	540.562	1018.321	21.773
d) $\tilde{\vartheta}_i^{S}(\hat{\sigma}_u^2, \hat{\rho})$	5581.765	3633.627	608.678	653.811

Figure 1. Map of Rathbun Lake watershed with (a-b-c) EBLUP and (d) Spatial EBLUP estimates for average erosion (Ton).



4.2. The average production of olives per farm in 42 LESs of the Tuscany region (Italy)

4.2.1. The survey data

The Farm Structure Survey is carried out once every two years and collects information on farm lands by type of cultivation, the amount of breeding, the kind of production, the structure and the amount of farm employment. The sample is selected by a stratified one stage design with self representation of larger farms (agricultural holdings). The sample size is 55,030 farms: 52,713 of them are drawn from the 2,150,000 firms of the so-called European Community target, while the additional 2,317 are selected from the 440,000 firms of the so-called Italian target. The stratification is done in three phases. In the first stage, the 6,972 self-represented farms are included in the sample on the basis of their economic dimension and/or their utilized surface area and/or number of bovines. In the second stage, the residual EC targeted farms are divided into 407 strata utilizing dimensional, geographical and gross income parameters. Finally, the farms of the Italian target are stratified into 21 regional strata. The optimal allocation of sample size to the strata is obtained minimizing the sampling error at regional and national level. Accurate estimates at sub-regional level require either the enlargement of the sample in provinces or municipalities or by the application of small area estimation models.

4.2.2. Estimates at LESs level

The Tuscany region is divided in 42 LES. The objective of inference is the average production of olives per farm ($\vartheta = \bar{y}$) measured in quintals for each of the 42 small areas (LES). Auxiliary data at the LES level were available for this study. The following models have been applied to estimation at LES level:

a) EBLUP ($\tilde{\vartheta}_i^{th}(\hat{\sigma}_u^2)$) with utilized surface area for olive production (ha);
b) EBLUP ($\tilde{\vartheta}_i^{th}(\hat{\sigma}_u^2)$) with geographical coordinates (in the UTM coordinate system) of the centroid of each LES plus utilized surface area for olive production (ha);
c) EBLUP ($\tilde{\vartheta}_i^{th}(\hat{\sigma}_u^2)$) with three territorial variables (the percentage of hill, mountain and plain for each LES) plus utilized surface area for olive production (ha);
d) EBLUP ($\tilde{\vartheta}_i^{th}(\hat{\sigma}_u^2)$) with geographical coordinates of the centroid of each LES and three territorial variables (the percentage of hill, mountain and plain for each LES) plus utilized surface area for olive production (ha);
e) Spatial EBLUP $\tilde{\vartheta}_i^{S}(\hat{\sigma}_u^2, \hat{\rho})$ with utilized surface area for olive production (ha);

f) Spatial EBLUP $\tilde{\vartheta}_i^{S}(\hat{\sigma}_u^2, \hat{\rho})$ with three territorial variables (the percentage of hill, mountain and plain for each LES) plus utilized surface area for olive production (ha);

The neighbourhood structure *W* is defined as follows: spatial weight, w_{ij} , is 1 if area *i* shares an edge with area *j* and 0 otherwise. For an easier interpretation, the general spatial weight matrix is defined in row standardized form, in which the row elements sum to one. Again, sampling variances, φ_i are estimated smoothing the sampling error associated to the population level estimator (Rao, 1998). The value of the estimated spatial autoregressive coefficient $\hat{\rho}$ is 0.859 ($s.e. = 0.113$) with the ML procedure for model (e) and suggests appreciable strength of spatial relationship. Under model (f) the spatial autoregressive coefficient $\hat{\rho}$ is 0.862 ($\mathbf{s.e.} = \mathbf{0.086}$).

Figure 2 displays the maps of Tuscany with (a-b-c-d) EBLUP and (e-f) Spatial EBLUP estimates for average production of olives per farm for each of the 42 LES.

Table 3 shows the correlations between the estimates obtained under models (a), (b), (c), (d), (e), (f). It can be noted that all the correlations are high and positive. This confirms that all the estimates are concordant, even if the introduction of geographical information reduces the smoothing in maps (b), (c), (d). The SEBLUP estimator allows for a clearer identification of southern LESs where the average production per farm is higher. Unfortunately the Census 2000 does not provide data about productions, but previous knowledge about local economies in southern Tuscany confirms our results (ISTAT, 2005). Table 4 reports the average estimated MSE per farm and its decomposition in g_1, g_2 and g_3 for the EBLUP and Spatial EBLUP estimators.

Two things stand out from Table 4: first, the introduction of geographic information improves the estimates obtained by EBLUP and SEBLUP by reducing the MSE; second, the SEBLUP estimator with model (f) has the best performance in terms of estimated MSE.

Under models (e) and (f) the value of the estimated spatial autoregressive coefficient $\hat{\rho}$ is 0.859 ($s.e. = 0.113$) with the ML procedure for model (e) and $\hat{\rho}$ is 0.862 ($\mathbf{s.e.} = \mathbf{0.086}$) even when including territorial variables in the fixed part of the model (f): this suggests appreciable strength of spatial relationship.

The components g_1, g_2 and g_3 for EBLUP and Spatial EBLUP estimators behave as we expected: g_1 is the largest one in all the models, g_3 in models (e), (f) is larger in comparison with the EBLUP models, due to the estimation of the additional parameter ρ.

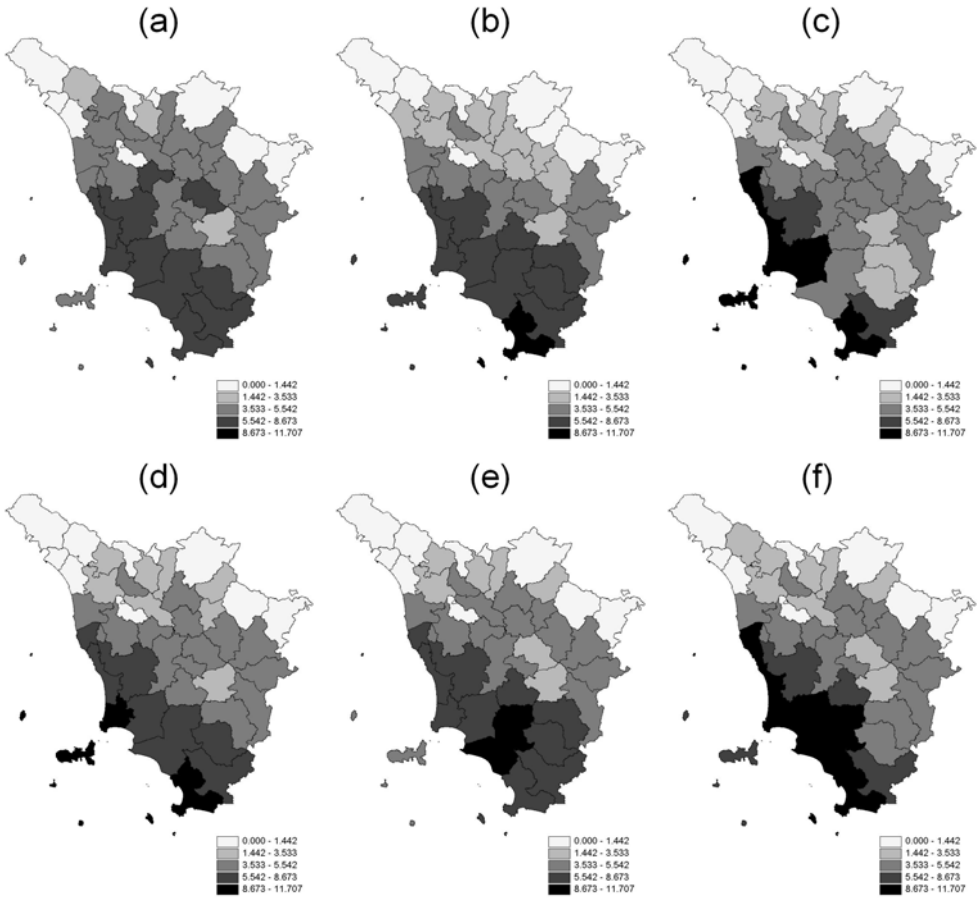
Table 3. Correlation matrix between the small area estimates of the average production of olives per farm.

	$\tilde{\vartheta}_i^{fh}(\hat{\sigma}_u^2)$	$\tilde{\vartheta}_i^{fh}(\hat{\sigma}_u^2)$ *+geographic coordinates*	$\tilde{\vartheta}_i^{fh}(\hat{\sigma}_u^2)$ *+territorial variables*	$\tilde{\vartheta}_i^{fh}(\hat{\sigma}_u^2)$ *+geographic coordinates + territorial variables*	$\tilde{\vartheta}_i^{S}(\hat{\sigma}_u^2$	$\tilde{\vartheta}_i^{S}(\hat{\sigma}_u^2,\hat{\rho})$ *+ territorial variables*
$\tilde{\vartheta}_i^{fh}(\hat{\sigma}_u^2)$	1					
$\tilde{\vartheta}_i^{fh}(\hat{\sigma}_u^2)$ *+geographic coordinates*	0.86	1				
$\tilde{\vartheta}_i^{fh}(\hat{\sigma}_u^2)$ *+ territorial variables*	0.75	0.82	1			
$\tilde{\vartheta}_i^{fh}(\hat{\sigma}_u^2)$ *+geographic coordinates+ territorial variables*	0.81	0.96	0.94	1		
$\tilde{\vartheta}_i^{S}(\hat{\sigma}_u^2,\hat{\rho})$	0.91	0.89	0.73	0.83	1	
$\tilde{\vartheta}_i^{S}(\hat{\sigma}_u^2,\hat{\rho})$ *+ territorial variables*	0.82	0.88	0.94	0.94	0.89	1

Table 4. Average Estimated MSE of EBLUP and Spatial EBLUP estimators

Estimator	mse	g_1	g_2	g_3
a) $\tilde{\vartheta}_i^{fh}(\hat{\sigma}_u^2)$	3.238	2.983	0.162	0.044
b) $\tilde{\vartheta}_i^{fh}(\hat{\sigma}_u^2)$ *+geographic coordinates*	1.516	1.223	0.246	0.021
c) $\tilde{\vartheta}_i^{fh}(\hat{\sigma}_u^2)$ *+territorial variables*	1.643	1.141	0.454	0.020
d) $\tilde{\vartheta}_i^{fh}(\hat{\sigma}_u^2)$ *+geographic coordinates+territorial variables*	1.429	0.804	0.586	0.015
e) $\tilde{\vartheta}_i^{S}(\hat{\sigma}_u^2,\hat{\rho})$	2.367	2.117	0.125	0.061
f) $\tilde{\vartheta}_i^{S}(\hat{\sigma}_u^2,\hat{\rho})$ *+territorial variables*	1.407	0.934	0.409	0.029

Figure 2. Map of Tuscany with (a-b-c-d) EBLUP and (e-f) Spatial EBLUP estimates for average production of olives per farm (quintal)



5. Final remarks

In this paper the investigation has been focused on alternative specification of the models underlying the EBLUP and SEBLUP predictors in order to incorporate geographic information into the model-based estimation of the small area parameters. Our point is that in the presence of spatial dependence, better estimates can be obtained by using the spatial information both in the fixed part of the models and in the random part, even by specifying models with spatially correlated random area effects.

The evidence from the case studies is that Spatial EBLUP, with correlated random area effects following a SAR process performs better when the spatial correlation in the study variable is high. However, the inclusion of covariates that

capture the spatial effects may be useful even when the strength of spatial link is weak.

The methodology presented allows us to make use of all the informative components of the survey data including the geographical ones. This is a relevant opportunity for environmental and agro-environmental studies where geographic information plays a fundamental role for a better understanding of the spatial pattern of the phenomena under analysis.

The spatial approach presented here is not free from limitations. Both EBLUP and SEBLUP are variable specific solutions: it is a matter of fact that geographic information relevant for a study variable can not be relevant for another. Nevertheless even if geographic information is not informative by itself, we have to accept that the spatial conformation of a study area (land use, elevation, percentage of hill, mountain and plain) are likely to influence deeply many environmental and socio-economic phenomena and their distribution by small area of interest.

In addition, the findings are sensitive to the definition of the geographical units under analysis. The Modifiable Areal Unit Problem (MAUP - Unwin, 1996) is a potential source of error that can affect spatial studies which utilize aggregate data sources and also the small area estimation results. A simple strategy to deal with the problem of MAUP in small area estimation is to undertake analysis at multiple scales or zones. This can conflict with budget and time constraints which often condition the small area statistics production process.

Furthermore, there is no unique way of defining the matrix of spatial interaction, and the results may be sensitive to the choice of spatial interaction matrix. Further work is needed to explore the performance of Spatial EBLUP using more complex spatial contiguity matrices.

REFERENCES

BAILEY T.C., GATRELL A.C. (1995), *Interactive Spatial Data Analysis*, Longman, London.

CLIFF A.D., ORD J.K. (1981), *Spatial Processes. Models & Applications*, Pion Limited, London.

CRESSIE, N. (1991): Small-Area Prediction of Undercount Using the General Linear Model, *Proceedings of the Statistic Symposium 90: Measurement and Improvement of Data Quality*, Ottawa: Statistics Canada, 93—105.

FAY, R.E., HERRIOT, R.A. (1979), Estimates of income for small places: an application of James-Stein procedures to census data, *Journal of the American Statistical Association*, *74*, 269—277.

ISTAT. (2005), Rapporto annuale: la situazione del Paese nel 2004, *ISTAT — Centro diffusione dati*, Roma.

KACKAR R.N., HARVILLE D.A. (1984), Approximations for standard errors of estimators for fixed and random effects in mixed models. *Journal of the American Statistical Association*, *79*, 853—862.

OPSOMER J.D., BOTTS C., KIM J.Y. (2003), Small Area Estimation in Watershed Erosion Assessment Survey, *J. of the Agricultural, Biological, and Environmental Statistics,* 8, 2, 139—152.

PETRUCCI A., SALVATI N. (2005), Small Area Estimation for Spatial Correlation in Watershed Erosion Assessment, forthcoming in *Journal of Agricultural , Biological, and Environmental Statistics.*

PFEFFERMANN D. (2002), Small Area Estimation — New Developments and Directions, *International Statistical Review*, *70*, *1*, 125—143.

PRATESI M., SALVATI N. (2005), Small Area Estimation: the EBLUP estimator with autoregressive random area effects, *Report n° 261, Dipartimento di Statistica e Matematica Applicata all'Economia, Pisa.*

PRASAD N., RAO J.N.K. (1990), The Estimation of the Mean Squared Error of Small-Area Estimators, *Journal of the American Statistical Association*, *85*, *409*, 163—171.

RAO J.N.K. (1998), Small area estimation, *Encyclopedia of Statistical Sciences*, *2*, 621—628.

RAO J.N.K. (2003), *Small area estimation,* John Wiley & Sons, New York.

SAEI, A., CHAMBERS, R. (2003), Small Area Estimation under linear and generalized linear mixed models with time and area effects, *Southampton Statistical Sciences Research Institute*, *WP M03/15*, Southampton.

TOBLER, W. R. (1970) A computer movie simulating urban growth in the Detroit region, *Economic Geography*, 46, 234—40.

UNWIN, D. J. (1996), GIS, spatial analysis and spatial statistics, *Progress in Human Geography 20(4)*: 540—441.